

Variable Selection with Exponential Weights and ℓ_0 -Penalization

Ery Arias-Castro and Karim Lounici

University of California, San Diego and Georgia Institute of Technology

Abstract

In the context of a linear model with a sparse coefficient vector, exponential weights methods have been shown to achieve oracle inequalities for prediction. We show that such methods also succeed at variable selection and estimation under the necessary identifiability condition on the design matrix, instead of much stronger assumptions required by other methods such as the Lasso or the Dantzig Selector. The same analysis yields consistency results for Bayesian methods and BIC-type variable selection under similar conditions.

Keywords: Variable selection, model selection, sparse linear model, exponential weights, Gibbs sampler, identifiability condition.

1 Introduction

Consider the standard linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_\star + \mathbf{z}, \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^n$ is the response vector; $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the regression (or design) matrix, assumed to have normalized columns; $\boldsymbol{\beta}_\star \in \mathbb{R}^p$ is the coefficient vector; and $\mathbf{z} \in \mathbb{R}^n$ is white Gaussian noise, i.e., $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. As in general the model (1) is not identifiable, we let $\boldsymbol{\beta}_\star$ denote one of the coefficient vectors such that $\mathbf{X}\boldsymbol{\beta} = \mathbb{E}(\mathbf{y})$ of minimal support size. Then J_\star and s_\star denote the support and support size of $\boldsymbol{\beta}_\star$. We are most interested in the case where the coefficient vector is sparse, meaning s_\star is much smaller than p . As usual, we want to perform inference based on the design matrix \mathbf{X} and the response vector \mathbf{y} . The three main inference problems are:

- *Prediction:* estimate the mean response vector $\mathbf{X}\boldsymbol{\beta}_\star$;
- *Estimation:* estimate the coefficient vector $\boldsymbol{\beta}_\star$;
- *Support recovery:* estimate the support J_\star .

These problems are not always differentiated and often referred to jointly as *variable/model selection* in the statistics literature, and *feature selection* in the machine learning literature. Being central to statistics, a large number of papers address these problems. We review the literature with particular emphasis on papers that advanced the theory of model selection. For penalized regression, we find (Shao, 1997), who provides necessary conditions and sufficient conditions under which the AIC/Mallows' C_p criteria and the BIC criteria are consistent. For example, AIC/Mallows' C_p are consistent when there is a unique $\boldsymbol{\beta}$ such that $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, and this $\boldsymbol{\beta}$ has a support of fixed size as $n, p \rightarrow \infty$. Also, BIC is consistent when the dimension p is fixed and the model is

identifiable — a condition that appears to be missing in that paper. BIC was recently shown in (Chen and Chen, 2008) to be consistent when the model is identifiable, $p = O(n^a)$ with $a < 1/2$ and the true coefficient vector has a support of fixed size as $n, p \rightarrow \infty$. They also propose an extended BIC for when a is larger. Assuming the size of the support of β_* is known, Raskutti et al. (2009) establish *prediction* and *estimation* performance bounds for best subset selection, and obtain information bounds for these problems. Relaxing to the ℓ_1 -norm penalty, the Lasso and the closely related Dantzig Selector were shown to be consistent when the design matrix satisfies a restricted isometric property (RIP) or has column vectors with low coherence; see (Bickel et al., 2009; Bunea, 2008; Bunea et al., 2007; Candès and Plan, 2009; Candès and Tao, 2007; Lounici, 2008; Meinshausen and Yu, 2009; Zhao and Yu, 2006) among others. With a carefully chosen non-concave penalty, (Fan and Peng, 2004) shows that consistent variable selection is possible when $p = O(n^{1/3})$. This condition on p was weakened in the follow-up paper (Fan and Lv, 2011), though with an additional restriction on the coherence (Condition (16) there). The strongest results in that line of work seem to appear in (Zhang, 2010), which suggests a minimax concave penalty that leads to consistent variable selection under much weaker assumptions. The classical forward stepwise selection, also known as orthogonal matching pursuit, which is shown in (Cai and Wang, 2011) to enable variable selection under an assumption of low coherence on the design matrix. Screening was studied in (Fan and Lv, 2008) in the ultrahigh dimensional setting, assuming the design is random. A combination of screening and penalized regression is explored in (Ji and Jin, 2010; Jin et al., 2012), with asymptotic optimality when the Gram matrix $\mathbf{X}^\top \mathbf{X}/n$ is (mildly) sparse.

A distinct line of research is the implementation of ℓ_0 -penalized regression via exponential weights (Catoni, 2004; Dalalyan and Salmon, 2011; Dalalyan and Tsybakov, 2007; Giraud, 2007; Juditsky et al., 2008; Lounici, 2007; Yang, 2004). This methodology, which has precedents in the Bayesian literature on model selection (Chipman et al., 2001), has the potential of striking a good compromise between statistical accuracy and computational complexity. While computational tractability has only been demonstrated in simulations, a number of sharp statistical results exist for the *prediction* problem. In particular, (Alquier and Lounici, 2011; Rigollet and Tsybakov, 2011) propose exponential weights procedures that achieve sharp sparsity oracle inequalities with no assumptions of the design matrix \mathbf{X} . Note that there exists no result in the literature concerning the problems of estimation and support recovery with an exponential weights approach. For a recent survey of the exponential weights literature, see (Rigollet and Tsybakov, 2012).

Our contribution is the following. We establish performance bounds for the version of exponential weights studied in (Alquier and Lounici, 2011) for the three main inference problems of *prediction*, *estimation* and *support recovery*. The methodology developed in the present paper is new and brings novel and interesting results to the sparse regression literature. The main feature of this methodology is that it only requires comparatively almost minimum assumptions on the design matrix \mathbf{X} . In particular, for *estimation* and *support recovery*, the conditions are slightly stronger than identifiability. Moreover, when the size of support is known, the exponential weights method is consistent under the minimum identifiability condition as long as the nonzero coefficients are large enough, close in magnitude to what is required by any method, in particular matching the performance of best subset selection (Raskutti et al., 2009). See also (Candès and Davenport, 2011; Verzelen, 2012; Zhang, 2007). An important by-product of our analysis are consistency results for BIC-type methods, i.e., variable selection with ℓ_0 -penalty, under similar conditions, extending the results of Chen and Chen (2008).

The rest of the paper is organized as follows. In Section 2, we describe in detail the methodology and state the main results. We also state similar results for variable selection with ℓ_0 -penalty and for the Bayesian model selection method of (Chipman et al., 2001). In Section 3, we compare the results we obtained for exponential weights with those established for other methods, in particular

the Lasso and MC+. In Section 4, we briefly discuss the algorithmic implementation of exponential weights, and show the result of some simple numerical experiments comparing exponential weights with other popular variable selection techniques in the literature. In Section 5, we discuss our results in the light of recent information bounds for model selection. The proofs of our main results are in Section 6.

2 Main results

We consider the version of exponential weights studied in (Alquier and Lounici, 2011), shown there to enjoy optimal oracle performance for the prediction problem. The procedure puts a sparsity prior on the coefficient vector and selects the estimates using the posterior distribution. We obtain a new *prediction* performance bound which is based on balancing the sparsity level and the size of the least squares residuals. The result does not assume any conditions on the design matrix. The task of *support recovery*, to be amenable, necessitates additional assumptions. We show that under near-identifiability conditions on the design matrix, the posterior concentrates on the correct subset of nonzero components with overwhelming probability, provided that these coefficients are sufficiently large — somewhat larger than the noise level. This immediately implies that the maximum a posteriori (MAP) is consistent. We then derive *estimation* performance guarantees in Euclidean norm and l_∞ -norm for the maximum a posteriori and posterior mean.

Throughout, we assume the noise variance σ^2 is known. We also assume that $p \geq n$ and remark that similar results hold when $n \geq p$, with p replaced by n in the bounds.

We use some standard notation. For any $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_d)^\top \in \mathbb{R}^d$ with $d \geq 1$ and $q \geq 1$, we define

$$\|\mathbf{u}\|_q = \left(\sum_{j=1}^d \|\mathbf{u}_j\|^q \right)^{1/q}, \quad \|\mathbf{u}\|_\infty = \max_{1 \leq j \leq d} \|\mathbf{u}_j\|.$$

Without loss of generality, we assume from now on that the predictors are normalized in the sense that

$$\frac{1}{\sqrt{n}} \|\mathbf{X}_j\|_2 = 1, \text{ for all } 1 \leq j \leq p. \quad (2)$$

For a subset $J \subset [p] := \{1, \dots, p\}$, let $\mathbf{X}_J = [\mathbf{X}_j, j \in J] \in \mathbb{R}^{n \times |J|}$, where \mathbf{X}_j denotes the j th column vector of \mathbf{X} . For a subset $J \subset [p]$, let M_J be the linear span of $\{\mathbf{X}_j, j \in J\}$ and let \mathbf{P}_J be the orthogonal projection onto M_J . Then, $\mathbf{P}_J^\perp := \mathbf{I}_n - \mathbf{P}_J$ is the orthogonal projection onto M_J^\perp . We say that a vector is s -sparse if its support is of size s .

2.1 Exponential weights

We start with the definition of a sparsity prior on the subsets of $[p]$, which favors subsets with small support. This leads to a pseudo-posterior, which is used in turn to define various exponential weights estimators.

- *The prior π .* Fix an upper bound $\bar{s} \geq 1$ on the support size, and a sparsity parameter $\lambda > 0$. The prior chooses the subset $J \subset [p]$ with probability

$$\pi(J) \propto \binom{p}{|J|}^{-1} e^{-\lambda|J|} \mathbb{I}_{\{|J| \leq \bar{s}\}}. \quad (3)$$

- *The posterior Π .* Given that the noise is assumed i.i.d. Gaussian with variance σ^2 , given a subset of variables $J \subset [p]$, the coefficient vector that maximizes the likelihood is the least squares estimate $\hat{\beta}_J$ with a maximum proportional to $\exp(-\|\mathbf{P}_J^\perp(\mathbf{y})\|_2^2/(2\sigma^2))$. In light of this, we define the following pseudo-posterior, which chooses $J \subset [p]$ with probability

$$\Pi(J) \propto \pi(J) \exp\left(-\frac{\|\mathbf{P}_J^\perp(\mathbf{y})\|_2^2}{2\sigma^2}\right). \quad (4)$$

The prior π enforces sparsity and focuses on subsets of size not exceeding \bar{s} . Without additional knowledge, we shall take $\bar{s} = p$. The exponential factor in $\|\mathbf{P}_J^\perp(\mathbf{y})\|_2^2$ in the posterior enforces fidelity to the observations. Note that Π is not a true posterior because no prior is assumed for β_\star ; we elaborate on this point in Section 2.4. The variance term $2\sigma^2$ corresponds to the temperature T in a standard Gibbs distribution. We will calibrate the procedure via the sparsity exponent λ in (3), though we could have done so via the temperature as well. Remember that we assume that σ^2 is known. When the variance is unknown, we can replace it with a consistent estimator $\hat{\sigma}^2$.

Based on the pseudo-prior Π , it is natural to consider the maximum a posteriori (MAP) support estimate, defined as

$$\hat{J}_{\text{map}} = \arg \max_J \Pi(J). \quad (5)$$

This leads to considering the MAP coefficient estimate. For any $J \subset [p]$, let $\hat{\beta}_J$ denote the least squares coefficient vector for the sub-model $(\mathbf{X}_J, \mathbf{y})$ with minimum Euclidean norm — so that $\hat{\beta}_J$ is unique even when the columns of \mathbf{X}_J are linearly dependent. When the columns of \mathbf{X}_J are linearly independent, the standard formula applies

$$\hat{\beta}_J = (\mathbf{X}_J^\top \mathbf{X}_J)^{-1} \mathbf{X}_J^\top \mathbf{y}. \quad (6)$$

Note that $\mathbf{I}(s)$ guarantees that (6) holds when $|J| \leq s$. The MAP coefficient estimate is then defined as $\hat{\beta}_{\text{map}} = \hat{\beta}_{\hat{J}_{\text{map}}}$.

We found that the MAP is not as stable as the posterior mean

$$\hat{\beta}_{\text{mean}} = \sum_J \Pi(J) \hat{\beta}_J. \quad (7)$$

We establish results for both of them.

2.2 Prediction

We establish a new sparsity oracle inequality for the prediction problem. We show that, in terms of prediction performance, the maximum a posteriori and posterior mean come within a log factor of that of the oracle estimator $\hat{\beta}_{J_\star}$:

$$\|\mathbf{X} \hat{\beta}_{J_\star} - \mathbf{X} \beta_\star\|_2 = \|\mathbf{P}_{J_\star} \mathbf{z}\|_2 = O_P(\sigma \sqrt{s_\star}).$$

Theorem 1. *Consider a design matrix \mathbf{X} with $p \geq n$ and normalized column vectors (2). Assume $\lambda = (62 + 12c) \log p$ for some $c > 0$. Then with probability at least $1 - p^{-c}$,*

$$\|\mathbf{X} \hat{\beta}_{\text{map}} - \mathbf{X} \beta_\star\|_2 \leq \sigma \sqrt{8s_\star \lambda} \quad \text{and} \quad \|\mathbf{X} \hat{\beta}_{\text{mean}} - \mathbf{X} \beta_\star\|_2 \leq \sigma \sqrt{12s_\star \lambda}. \quad (8)$$

Note that here, and anywhere else in the paper, what is true of $\hat{\beta}_{\text{map}}$ is true of $\hat{\beta}_J$ for any J such that $\Pi(J) \geq \Pi(J_\star)$.

In (Alquier and Lounici, 2011), a similar sparsity oracle inequality is established in expectation using the approach by Stein's Lemma from (Leung and Barron, 2006). Here, we use instead the concentration property of the posterior Π and show that the oracle inequality also holds true in probability. Note that Alquier and Lounici (2011) also established an oracle inequality in probability for a different exponential weights procedure that requires the knowledge of $\|\beta_\star\|_1$. Our result constitutes an improvement since we do not require such knowledge.

2.3 Concentration of the posterior and support recovery

Our performance bounds for support recovery rely, as they should, on concentration properties of the posterior Π . We first prove that, without any condition on the design matrix \mathbf{X} , the posterior Π concentrates on subsets of small size.

Proposition 1. *Consider a design matrix \mathbf{X} with $p \geq n$ and normalized column vectors (2). For some $\varepsilon > 0$ and $c \geq 1$, take*

$$\lambda = \frac{1 + \varepsilon}{\varepsilon} (23 + 5c) \log p. \quad (9)$$

Then, with probability at least $1 - 2p^{-c}$, $\Pi(J) < \Pi(J_\star)$ for all $J \subset [p]$ such that $|J| > (1 + \varepsilon)s_\star$, and in fact

$$\Pi(J : |J| > (1 + \varepsilon)s_\star) \leq 4p^{-c} \Pi(J_\star). \quad (10)$$

2.3.1 Identifiability

Actual support recovery requires some additional conditions, the bare minimum being that the model is identifiable.

Condition $\mathbf{I}(s)$: For any subset $J \subset \{1, \dots, p\}$ of size $|J| \leq s$, the submatrix \mathbf{X}_J is full-rank.

This condition characterizes the identifiability of the model as stated in the following simple result.

Lemma 1. *Assuming $\beta_\star \in \mathbb{R}^p$ is s_\star -sparse, it is identifiable if, and only if, $\mathbf{I}(2s_\star)$ is satisfied.*

In this paper, we establish that exponential weights, and also ℓ_0 -penalized variable selection, allow for support recovery and estimation under the condition $\mathbf{I}((2 + \varepsilon)s_\star)$ for any $\varepsilon > 0$ fixed, as long as the non-zero entries of the coefficient vector are sufficiently large. In fact, $\mathbf{I}(2s_\star)$ suffices when s_\star is known.

While $\mathbf{I}(s)$ is qualitative, results on estimation and support recovery necessarily require a quantitative measure of correlation in the covariates. The following quantity appears in the performance bounds we derive for exponential weights and related methods: for any integer $s \geq 1$, define

$$\nu_s = \min_{J \subset [p] : |J| \leq s} \min_{\mathbf{u} \in \mathbb{R}^{|J|} : \|\mathbf{u}\|_2 = 1} \frac{1}{\sqrt{n}} \|\mathbf{X}_J \mathbf{u}\|_2. \quad (11)$$

Equivalently, ν_s is the smallest singular value of among submatrices of $\frac{1}{\sqrt{n}} \mathbf{X}$ made of at most s columns. Note that, indeed, $\mathbf{I}(s)$ is equivalent to $\nu_s > 0$.

2.3.2 Support recovery

We now state the main result concerning the support recovery problem. It states that, under $\mathbf{I}((2 + \varepsilon)s_\star)$, the posterior distribution Π concentrates sharply on the support of β_\star — which we assumed to be s_\star -sparse — as long as λ and the nonzero coefficients are sufficiently large.

Theorem 2. *Consider a design matrix \mathbf{X} , with $p \geq n$ and normalized column vectors (2), that satisfies Condition $\mathbf{I}((2 + \varepsilon)s_\star)$ for some fixed $\varepsilon > 0$. Assume that (9) holds and*

$$\min_{j \in J_\star} |\beta_{\star,j}| \geq \rho := \frac{3\sigma\sqrt{\lambda/n}}{\nu_{(2+\varepsilon)s_\star}}. \quad (12)$$

Then, with probability at least $1 - 2p^{-c}$, $\Pi(J_\star) > \Pi(J)$ for all J , and in fact

$$\Pi(J_\star) \geq 1 - 4p^{-c}.$$

Under the conditions of Theorem 2, some straightforward calculations imply that $\hat{J}_{\text{map}} = J_\star$ with probability at least $1 - 6p^{-c}$. In particular, as $p \rightarrow \infty$, the MAP consistently recovers the support of the coefficient vector. Note that the same is true for a subset drawn from Π .

The result applies in the ultra-high dimensional setting where p is exponential in n , as long as the conditions are met. Characterizing design matrices \mathbf{X} that satisfy $\mathbf{I}((2 + \varepsilon)s_\star)$ in the ultra-high dimensional setting is an interesting open question beyond the scope of this paper.

We mention that, if s_\star is known and we restrict the prior over subsets J of size exactly s_\star , then the same conclusions are valid with $\varepsilon = 0$ and $\nu_{(2+\varepsilon)s_\star}$ replaced by ν_{2s_\star} in (12), yielding consistent support recovery under the minimum identifiability condition $\mathbf{I}(2s_\star)$. In Section 3, we show that the Lasso estimator requires much more restrictive conditions on the design matrix and β_\star to ensure it selects the correct variables with high probability.

Finally, we note that the concentration is even stronger. Under the same conditions, if

$$\lambda = \frac{(1 + \varepsilon)(23 + 5c) + m}{\varepsilon} \log p,$$

then

$$\sum_{J \subset [p]: J \neq J_\star} |J|^m \Pi(J) \leq 4p^{-c} \Pi(J_\star). \quad (13)$$

We will use this refinement in the proof of Theorem 5.

2.3.3 Estimation

Armed with results for the support recovery and prediction problems, we establish corresponding bounds for the estimation problem. Our first result is a simple consequence of Theorem 1 and Proposition 1.

Theorem 3. *Consider a design matrix \mathbf{X} with $p \geq n$ and normalized column vectors (2). Assume λ satisfies (9) with $\varepsilon \leq 1/2$. Then with probability at least $1 - 3p^{-c}$, we have*

$$\|\hat{\beta}_{\text{map}} - \beta_\star\|_2 \leq \sigma \sqrt{\frac{8s_\star\lambda}{n\nu_{(2+\varepsilon)s_\star}^2}}.$$

We continue with bounds on the estimation error, this time in terms of the l_∞ -norm. Based on Theorem 2 (and its proof), we deduce the following.

Theorem 4. *Let the conditions of Theorem 2 be satisfied. Then, with probability at least $1 - 7p^{-c}$, we have*

$$\|\hat{\beta}_{\text{map}} - \beta_{\star}\|_{\infty} \leq \sigma \sqrt{\frac{2(c+1) \log p}{n\nu_{s_{\star}}^2}}. \quad (14)$$

We emphasize that this estimator requires only the near minimum condition $\mathbf{I}((2+\varepsilon)s_{\star})$ and that the nonzero components of β_{\star} are somewhat larger than the noise level in (12) to achieve the optimal (up to logs) dependence on n, p of the l_{∞} -norm estimation bound. We will develop this point further in our comparison with the Lasso.

We now study the performances of the posterior mean $\hat{\beta}_{\text{mean}}$ and that of the following variant

$$\tilde{\beta} = \sum_{J \subset [p]: \nu_J > 0} \Pi(J) \hat{\beta}_J, \quad \nu_J := \min_{\mathbf{u} \in \mathbb{R}^{|J|}: \|\mathbf{u}\|_2=1} \frac{1}{\sqrt{n}} \|\mathbf{X}_J \mathbf{u}\|_2. \quad (15)$$

Define the quantity $\nu_{\min} = \min_{J \subset [p]: \nu_J > 0} \nu_J$, and note that $\nu_{\min} > 0$.

Theorem 5. *Let the conditions of Theorem 2 be satisfied and let $c \geq 1$.*

1. *Take $\lambda = \frac{(1+\varepsilon)(23+5c)+1}{\varepsilon} \log p$. Then, with probability at least $1 - 4p^{-c}$,*

$$\|\tilde{\beta} - \beta_{\star}\|_{\infty} \leq \sigma \sqrt{\frac{2(c+1) \log p}{n\nu_{s_{\star}}^2}} + \frac{3}{\nu_{\min} p^c} \left[\sigma \sqrt{(20+4c) \frac{\log p}{n}} + \frac{\|\mathbf{X} \beta_{\star}\|_2}{\sqrt{n}} + \nu_{\min} \|\beta_{\star}\|_{\infty} \right].$$

2. *If in addition $\mathbf{I}(\bar{s})$ is satisfied,*

$$\|\hat{\beta}_{\text{mean}} - \beta_{\star}\|_{\infty} \leq \sigma \sqrt{\frac{2(c+1) \log p}{n\nu_{s_{\star}}^2}} + \frac{3}{\nu_{\bar{s}} p^c} \left[\sigma \sqrt{(20+4c) \frac{\log p}{n}} + \frac{\|\mathbf{X} \beta_{\star}\|_2}{\sqrt{n}} + \nu_{\bar{s}} \|\beta_{\star}\|_{\infty} \right].$$

3. *If in addition $\mathbf{I}(s_{\star} + \bar{s})$ is satisfied and $\lambda \geq (62+4c) \log p$,*

$$\|\hat{\beta}_{\text{mean}} - \beta_{\star}\|_{\infty} \leq \sigma \sqrt{\frac{2(c+1) \log p}{n\nu_{s_{\star}}^2}} + \frac{2\sqrt{10}\sigma}{\sqrt{n}\nu_{s_{\star}+\bar{s}}} \left[\frac{2\sqrt{s_{\star}}}{p^c} + \frac{1}{p^{s_{\star}}} \right]. \quad (16)$$

We note that $\hat{\beta}_{\text{mean}}$ requires at least $\mathbf{I}(\bar{s})$. (Recall that we assume \bar{s} is known such that $s_{\star} \leq \bar{s}$.) In practice, when the sparsity is unknown, we make a conservative choice $\bar{s} \gg 2s_{\star}$ so that $\mathbf{I}(\bar{s})$ is substantially more restrictive than $\mathbf{I}(2s_{\star})$. Typically, we assume that $s_{\star} = O\left(\frac{n}{\log p}\right)$ and we take \bar{s} of this order of magnitude. We will see below in Section 2.3.4 that for Gaussian design, the condition $\mathbf{I}(s_{\star} + \bar{s})$ is satisfied with probability close to 1. On the other hand, the estimation result for $\tilde{\beta}$ holds true under the near minimum condition $\mathbf{I}((2+\varepsilon)s_{\star})$. For both estimators, their estimation bounds depend on the quantities ν_{\min} , $\nu_{\bar{s}}$, $\|\mathbf{X} \beta_{\star}\|_2$ and $\|\beta_{\star}\|_{\infty}$ which can potentially yield a sub-optimal rate of estimation. Note however the presence of the factor p^{-c} in the bound. In particular, if the nonzero components of β_{\star} are sufficiently large, then the quantities ν_{\min} , $\nu_{\bar{s}}$, $\|\mathbf{X} \beta_{\star}\|_2$ and $\|\beta_{\star}\|_{\infty}$ may be completely cancelled for a sufficiently large $c > 0$. If $\mathbf{I}(s_{\star} + \bar{s})$ is satisfied, then we can derive a bound that no longer depends on $\|\mathbf{X} \beta_{\star}\|_2$ and $\|\beta_{\star}\|_{\infty}$. We will also see below that this bound yields the optimal rate of l_{∞} -norm estimation (up to logs) for the estimator $\hat{\beta}_{\text{mean}}$ when the design matrix is Gaussian. Optimality considerations are further discussed in Section 5 based on recent information bounds obtained elsewhere.

2.3.4 Example: Gaussian design

The quintessential example is that of a random Gaussian design, where the *row* vectors of \mathbf{X} , denoted $\mathbf{x}_1, \dots, \mathbf{x}_n$, are independent Gaussian vectors in \mathbb{R}^p with zero mean and $p \times p$ covariance matrix Σ . If we assume that Σ has 1's on the diagonal, the resulting (random) design is just slightly outside our setting, since the columns vectors are not strictly normalized. Our results apply nevertheless. Therefore, it is of interest to lower-bound ν_s for such a design.

We start by relating \mathbf{X} and Σ . Consider $J \subset [p]$, and let Σ_J denote the principal submatrix of Σ indexed by J . By (Vershynin, 2010, Cor. 1.50 and Rem. 1.51), there is a numeric constant $C > 0$ such that, when $n \geq C|J|/\eta^2$, with probability at least $1 - 2\exp(-\eta^2 n/C)$, we have

$$\left\| \frac{1}{n} \mathbf{X}_J^\top \mathbf{X}_J - \Sigma_J \right\| \leq \eta \|\Sigma_J\|,$$

where $\|\cdot\|$ denotes the matrix spectral norm. When this is the case, by Weyl's theorem (Stewart and Sun, 1990, Cor. IV.4.9),

$$\lambda_{\min} \left(\frac{1}{n} \mathbf{X}_J^\top \mathbf{X}_J \right) \geq \lambda_{\min}(\Sigma_J) - \eta \lambda_{\max}(\Sigma_J),$$

where $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalues of a symmetric matrix \mathbf{A} . Define

$$\eta_{\Sigma}(s) = \max_{J: |J| \leq s} \frac{\lambda_{\max}(\Sigma_J)}{\lambda_{\min}(\Sigma_J)}, \quad \lambda_{\Sigma}(s) = \min_{J: |J| \leq s} \lambda_{\min}(\Sigma_J).$$

Assume that

$$n \geq \frac{aCs \log p}{\eta_{\Sigma}(s)^2},$$

for some $a \geq 2$. Then, with probability at least $1 - 2p^{-a/2}$,

$$\nu_s \geq \frac{\lambda_{\Sigma}(s)^{1/2}}{2}.$$

For example, in standard compressive sensing where Σ is the identity matrix, we have $\eta_{\Sigma}(s) = \lambda_{\Sigma}(s) = 1$ for all s , in which case with high probability $\nu_s \geq 1/2$ when $n \geq 2Cs \log p$. Consequently, the l_{∞} -norm estimation bounds in (14) and (16) are of the order $b\sigma \sqrt{\log(p)/n}$ for some numerical constant $b > 0$. Again, the constants are loose in this discussion.

2.4 Bayesian variable selection with an independence prior

Many Bayesian techniques for model selection have proposed in the literature; see (Chipman et al., 2001) for a comprehensive review. That same paper suggests a procedure similar to ours, except that it is a bonafide Bayesian model and they use the following independence sparsity prior

$$\tilde{\pi}(J) = \omega^{|J|} (1 - \omega)^{p-|J|},$$

where $\omega \in (0, 1)$ controls the sparsity level. Roughly, λ for our prior corresponds to $\log(1 - 1/\omega)$ for this prior. It so happens that, the same arguments lead to the same results. Also, as argued in (Chipman et al., 2001, Sec. 3.3), the fully-specified Bayesian model with prior $\beta_J \sim \mathcal{N}(\mathbf{0}, (\mathbf{X}_J^\top \mathbf{X}_J)^{-1})$ is very closely related to our exponential weights method.

2.5 ℓ_0 -penalized variable selection

Chen and Chen (2008) not only showed that BIC was consistent when $p < \sqrt{n}$ (under some mild conditions on the design matrix), they also suggested a modification of the penalty term to yield a method that is consistent for larger values of p when the number of variables in the true (i.e., sparsest) model s_* is bounded independently of n or p .

By a simple modification of our arguments, our results for the exponential weights MAP is seen to apply to

$$\hat{J} = \arg \min_{J: |J| \leq \bar{s}} \mathbf{y}^\top (\mathbf{I} - \mathbf{P}_J) \mathbf{y} + \lambda |J|.$$

Consequently, our work extends that of Chen and Chen (2008) to the case where s_* increases with p .

3 Comparison with the Lasso and MC+

In this section, we compare the theoretical performances of our procedure with other well-known l_∞ -estimation and support recovery techniques used in high-dimensional variable selection.

3.1 Lasso

The Lasso estimator is the solution of the convex minimization problem

$$\hat{\boldsymbol{\beta}}^L = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + 2\lambda_L \|\boldsymbol{\beta}\|_1 \right\},$$

where $\lambda_L = A\sigma\sqrt{\log(p)/n}$, $A > 0$ and $\|\cdot\|_1$ is the l_1 -norm. The Lasso has received considerable attention in the literature over the last few years (Bach, 2008; Bunea, 2008; Bunea et al., 2007; Meinshausen et al., 2006; Meinshausen and Yu, 2009; Zhao and Yu, 2006). It is not our goal to make here an exhaustive presentation of all existing results. We refer to Chapter 4 in (Lounici, 2009) and the references cited therein for a comprehensive overview of the literature.

Concerning the l_∞ -norm estimation and support recovery problems, the most popular assumption is the Irrepresentable Condition (Bach, 2008; Lounici, 2009; Meinshausen and Yu, 2009; Wainwright, 2006; Zhao and Yu, 2006) denoted from now on by $\mathbf{IC}(s_*)$. See for instance Assumption 4.2 in (Lounici, 2009). The condition $\mathbf{IC}(s_*)$ is strictly more restrictive than the identifiability $\mathbf{I}(2s_*)$ and does not hold true in general when the columns of the design matrix \mathbf{X} are not weakly correlated. Define $d_* = \|\Psi_*^{-1} \text{sign}(\boldsymbol{\beta}_*)\|_\infty$ where $\Psi_* := \frac{1}{n} \mathbf{X}_{J_*}^\top \mathbf{X}_{J_*}$. The following result is the key to our analysis. Let $\mathbf{IC}(s_*)$ holds true and let the nonzero components of $\boldsymbol{\beta}_*$ be sufficiently large: $\min_{j \in J_*} |\beta_{*,j}| > A\sigma d_* \sqrt{\log(p)/n}$. Then, with probability at least $1 - 2p^{1-\frac{A^2}{16}} - s_* p^{-\frac{A^2}{2}}$, the Lasso solution is unique and satisfies

$$c\sigma d_* \sqrt{\frac{\log p}{n}} \leq \|\hat{\boldsymbol{\beta}}^L - \boldsymbol{\beta}_*\|_\infty \leq C\sigma d_* \sqrt{\frac{\log p}{n}}, \quad (17)$$

for some numerical constants $C \geq c > 0$ that can depend only on A . See Theorem 4.1 in (Lounici, 2009) for a more precise statement.

We say that a l_∞ -norm estimation rate is optimal if it is of the form $\alpha\sigma\sqrt{\log(p)/n}$ where $\alpha > 0$ is an absolute constant as in the case of gaussian sequence model ($n = p$ and $\mathbf{X} = \mathbf{I}_n$ the $n \times n$ identity matrix). In view of the previous display, the Lasso does not attain in most cases the optimal l_∞ -norm estimation rate. Indeed, the quantity d_* generally depends on s_* unless the correlations between the columns of the design matrix \mathbf{X} are very weak. Consider for the instance the case

where $\min_{j \neq k} |(\Psi_\star^{-1})_{j,k}| \geq \rho$ for some fixed $\rho > 0$. Then, we can easily find a s_\star -sparse vector β_\star such that $d_\star \geq \rho s_\star$ and the l_∞ -norm estimation rate of the Lasso is then suboptimal by a factor s_\star .

Unlike the Lasso, our exponential weights procedure does not suffer from this limitation. Indeed, our procedure achieves the optimal l_∞ -norm estimation rate and support recovery provided that Condition $\mathbf{I}((2 + \varepsilon)s_\star)$ holds true, which can be the case even for design matrices \mathbf{X} with strongly correlated columns.

Gaussian design. Consider the Gaussian design of Section 2.3.4, but assume now that $\Sigma = I_{p \times p}$ the $p \times p$ identity matrix. Although the design $\frac{1}{\sqrt{n}}\mathbf{X}$ satisfies the restricted isometry with probability close to 1, there is no guarantee that \mathbf{X} also satisfies an irrepresentable condition $\mathbf{IC}(s_\star)$. Let's assume that this is the case for the sake of comparison. Then, we can show with probability close to 1 that Ψ_\star^{-1} satisfies the mutual coherence condition $\max_{j \neq k} |(\Psi_\star^{-1})_{j,k}| \leq \frac{1}{\sqrt{s_\star}}$ where the dependence on s_\star cannot be improved. Thus, we get $d_\star \leq \sqrt{s_\star}$ and we cannot guarantee the optimality of the l_∞ -norm estimation bound for the Lasso under the irrepresentable condition. Consequently, we need the condition $\min_{j \in J_\star} |\beta_{\star,j}| \geq C\sigma\sqrt{s_\star \log(p)/n}$ for some absolute constant $C > 0$ in order to guarantee exact support recovery for the Lasso. This condition is to be compared to (12) for the exponential weights estimators. In that case, we have $\nu_{\tilde{s}} > 1/2$ with probability close to 1 when $\tilde{s} = O(n/\log p)$, so that (12) becomes simply $\min_{j \in J_\star} |\beta_{\star,j}| \geq C\sigma\sqrt{\log(p)/n}$ for some numerical constant $C > 0$. This condition is less restrictive than that for the Lasso by a factor $\sqrt{s_\star}$. Next, we note also that for a Gaussian design, the estimation bounds (14) and (16) for the exponential weights estimators are optimal (up to log) whereas the estimation bound for the Lasso contains the additional factor $\sqrt{s_\star}$.

Recently, in the framework of instrumental regression, [Gautier and Tsybakov \(2011\)](#) established for an l_1 -norm minimization procedure $\hat{\beta}^D$ close to the Dantzig selector (see (3.5) there) that with probability close to 1

$$\|D_{\mathbf{X}}^{-1}(\hat{\beta}^D - \beta_\star)\|_q \leq 2\sigma\sqrt{\frac{\log p}{\kappa_{q,J_\star}^2 n}},$$

where the sensitivity

$$\kappa_{q,J_\star} := \inf_{\Delta \in C_{J_\star} : \|\Delta\|_q = 1} \left| \frac{1}{n} D_{\mathbf{X}} \mathbf{X}^\top \mathbf{X} D_{\mathbf{X}} \Delta \right|,$$

with $C_{J_\star} = \left\{ \Delta \in \mathbb{R}^p : \|\Delta_{J_\star^c}\|_1 \leq \frac{1+c}{1-c} \|\Delta_{J_\star}\|_1 \right\}$ for some $0 < c < 1$, $D_{\mathbf{X}} = \text{diag}(\mathbf{X}_{1*}, \dots, \mathbf{X}_{p*})$, $\mathbf{X}_{k*} = \max_{1 \leq i \leq n} |\mathbf{X}_k^{(i)}|$ for any $1 \leq k \leq p$ and the $\mathbf{X}_k^{(i)}$ are the components of \mathbf{X}_k . An enticing property of the sensitivity approach is that the quantities $\kappa_{q,J}$ can be computed in reasonable time for small J , yielding a computationally tractable procedure to build confidence interval for the estimation of β_\star . The downside is that without any further conditions on \mathbf{X} , the optimal dependence of the bound on s_\star is not clear. For instance, assume in addition that \mathbf{X} satisfies a restricted eigenvalue condition as in ([Bickel et al., 2009](#)), then the dependence of the above bound on s_\star can be proved to be optimal for any $1 \leq q \leq 2$ (See Section 9 in [Gautier and Tsybakov \(2011\)](#)). However, if we only assume that the condition $\mathbf{I}((2 + \varepsilon)s_\star)$ is satisfied, then we cannot establish a clear comparison between the exponential weights and $\hat{\beta}^D$. The exponential weights estimator achieves the optimal estimation rate while the dependence on s_\star of the l_q -norm estimation bound is not explicit for $\hat{\beta}^D$.

3.2 MC+

The MC+ estimator initially proposed by [Zhang \(2010\)](#) is the solution of the following nonconvex minimization problem:

$$\hat{\beta}^{MC+} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \sum_{j=1}^p \Upsilon(|\beta_j|, \lambda_{MC}, \gamma) \right\}, \quad (18)$$

where $\lambda_{MC}, \gamma > 0$ and the MC+ penalty function Υ is nonconvex, equal to 0 outside a compact neighborhood of 0 and admits a nonzero right derivative at 0. See equations (2.1)-(2.3) in [\(Zhang, 2010\)](#) for more details.

The performance of this estimator is established in Theorem 1 of [\(Zhang, 2010\)](#), where the tuning of the parameter λ_{MC} requires the knowledge of s_\star (which is d° in that paper) and the optimal theoretical choice of γ is proportional to $\nu_{\bar{s}}^{-1}$ (which is d^* in that paper). Let us also emphasize that the choice $\gamma \propto \nu_{\bar{s}}^{-1}$ requires that $\mathbf{I}(\bar{s})$ is satisfied where \bar{s} is in practice a conservative upper bound on s_\star . In addition, this quantity $\nu_{\bar{s}}$ is delicate to compute in practice. Note that the exponential weights do not present the same limitations. Indeed, no prior knowledge of s_\star is required and we only need the condition $\mathbf{I}((2 + \varepsilon)s_\star)$ for an arbitrarily small $\varepsilon > 0$ to establish the consistency of $\hat{\beta}_{\text{map}}$ even if the parameter \bar{s} is chosen conservatively (for example, $\bar{s} = \lceil n/2 \rceil$ if no other information is available). In addition, the tuning of the parameters for the exponential weights do not require to compute any restricted eigenvalues. We mention that the assumptions in Theorem 1 of [\(Zhang, 2010\)](#) do not guarantee the identifiability of β_\star .

4 Numerical Experiments

In this section, we illustrate the performance of the procedure (4) in variable selection and l_∞ -norm estimation on a simulated data set. The posterior (4) is simulated via MCMC. In a nutshell, we construct an ergodic Markov chain $(\beta_t)_{t \geq 0}$ with invariant probability distribution the posterior (4). Then, we get from [\(Robert and Casella, 2004\)](#) that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=T_0+1}^{T_0+T} \beta_t = \hat{\beta}_{\text{mean}}, \quad \pi - a.s.,$$

where $T_0 \geq 0$ is an arbitrary number. In practice, we use $T_0 = 3000$ and $T = 7000$.

We refer to [\(Alquier and Lounici, 2011; Rigollet and Tsybakov, 2011\)](#) for more details on the computational aspect. Our numerical study follows those carried out in these references except that we concentrate on the l_∞ -norm estimation and variable selection performances of the procedure $\hat{\beta}_{\text{mean}}$. Note that the exponential weights procedures considered in the present paper and in [\(Alquier and Lounici, 2011; Rigollet and Tsybakov, 2011\)](#) differ only through the tuning of the parameters. [\(Alquier and Lounici, 2011; Rigollet and Tsybakov, 2011\)](#) consider indeed the prediction problem whereas we concentrate on the l_∞ -norm estimation and support recovery problems, which require a different tuning to guarantee the theoretical consistency. Note also that $\hat{\beta}_{\text{mean}}$ is not sparse since it is obtained as the expectation of the posterior Π . However, in view of Theorem 5, a simple thresholding of $\hat{\beta}_{\text{mean}}$ with a threshold of the order of the noise level yields consistent support recovery. In our simulations, we observe that few components of $\hat{\beta}_{\text{mean}}$ are significantly far from 0 whereas the remaining ones are extremely small thus making the choice of the threshold easy in practice. From now on, we will denote indifferently by AEW the procedure $\hat{\beta}_{\text{mean}}$ and the thresholded $\hat{\beta}_{\text{mean}}$.

Following the numerical experiments of (Candès and Tao, 2007), we consider the model (1) where \mathbf{X} is an $n \times p$ matrix with independent standard Gaussian entries, the target vector $\beta_\star = \mathbb{1}_{\{j \leq s_\star\}}$ for some fixed $s_\star \geq 1$ and the noise variance satisfies $\sigma^2 = \|\mathbf{X}\beta_\star\|^2/(9n)$. For each different setting of (n, p, s_\star) , we perform 100 replications of the model and compare our estimator AWE with other procedures in the literature on sparse estimation:

1. The Lasso estimator;
2. The MC+ estimator of (Zhang, 2010);
3. The SCAD estimator of (Fan and Li, 2001).

All these procedures are readily implemented in R. We use the glmnet package to compute the Lasso estimator and the ncvg package to compute the SCAD and MC+ estimators. The AWE estimator was computed through the MCMC algorithm described in Section 7 in (Rigollet and Tsybakov, 2011).

Figure 1 contains the comparative boxplots for the l_∞ -norm error over the 100 repetitions for the independent Gaussian design. Table 1 contains the average l_∞ -norm error and the standard deviation over the 100 repetitions for the independent Gaussian design. We observe that the AWE estimator outperforms the Lasso estimator and exhibit performances similar to MC+ and SCAD.

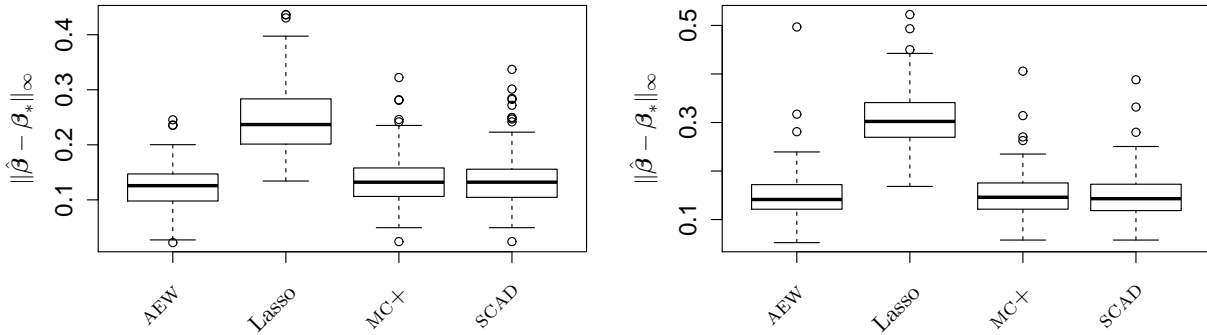


Figure 1: Independent Gaussian design. Boxplots of estimation performance measure $\|\hat{\beta} - \beta_\star\|_\infty$ over 100 realizations for the AEW, Lasso, MC+ and SCAD estimators. *Left:* $(n, p, s_\star) = (100, 200, 5)$. *Right:* $(n, p, s_\star) = (200, 1000, 10)$.

(n, p, s_\star)	AEW	Lasso	MC+	SCAD
$(100, 200, 5)$	0.124 (0.041)	0.249 (0.068)	0.137 (0.050)	0.138 (0.056)
$(200, 1000, 10)$	0.151 (0.055)	0.309 (0.063)	0.153 (0.051)	0.149 (0.050)

Table 1: Independent Gaussian design. Means and standard deviations of performance measures over 100 realizations for the AEW, Lasso, MC+ and SCAD estimators.

We note in our simulation study that the four procedures always select the s_\star active covariates but also select non-active ones. Table 2 contains the average support recovery false positive rate

over the 100 repetitions for the four procedures considered in this study. We observe that the Lasso tends to select too many covariates as was already known. The (thresholded) AWE estimator outperforms all other procedures in the support recovery problem.

(n, p, s_*)	AWE	Lasso	MC+	SCAD
$(100, 200, 5)$	1.60	21.55	1.75	3.02
$(200, 1000, 10)$	1.98	51.88	2.49	5.22

Table 2: Average support recovery false positive rate over 100 realizations for the AEW, Lasso, MC+ and SCAD estimators.

5 Discussion

We established some performance bounds for exponential weights when applied to solving the problems of *prediction*, *estimation* and *support recovery*, and deduced similar results for a slightly different Bayesian model selection procedure (Chipman et al., 2001) and ℓ_0 -penalized (BIC-type) variable selection. How sharp are these bounds? We did not optimize the numerical constants appearing in our results, simply because we believe our bounds are loose and also because there are no known sharp information bounds for these problems, except in specific cases (Jin et al., 2012). That said, there are some results available in the literature (Lounici et al., 2011; Raskutti et al., 2009; Verzelen, 2012) and our bounds come close to these. For example, from (Raskutti et al., 2009) we learn that, when $\mathbf{I}(2s_*)$ holds, there is a universal constant $C > 0$ such that, for any estimator $\hat{\beta}$ that knows s_* ,

$$\|\hat{\beta} - \beta_*\|_2 \geq C\sigma \sqrt{\frac{s_* \log(p/s_*)}{n\kappa_{2s_*}^2}}$$

with probability at least $1/2$, where

$$\kappa_s := \max_{J \subset [p]: |J| \leq s} \min_{\|\mathbf{u}\|=1} \frac{1}{\sqrt{n}} \|\mathbf{X}_J \mathbf{u}\|_2; \quad (19)$$

and from (Verzelen, 2012), we learn that, for another universal constant $C' > 0$,

$$\mathbb{E} \|\hat{\beta} - \beta_*\|_2^2 \geq C' \sigma^2 \left(\frac{s_* \log(ep/s_*)}{\kappa_{2s_*}^2} \vee \frac{1}{\nu_{2s_*}^2} \right).$$

Thus we see that our estimation bounds (14) and (16) come quite close to these information bounds. Of course, there is a trade-off with computational tractability, as computing the exponential weights estimates (of even approximating them) in polynomial time remains an open problem. That said, the numerical experiments show that these methods are promising.

6 Proofs

For the sake of brevity, we let $\|\cdot\| = \|\cdot\|_2$ throughout this section.

6.1 Proof of Theorem 1

Define $\xi_J = P_J(\mathbf{y}) - \mathbf{X}\beta_\star$. For $J \subset [p]$ with $|J| = s$, we have

$$\frac{\Pi(J)}{\Pi(J_\star)} = \frac{\binom{p}{s_\star}}{\binom{p}{s}} \exp\left(\lambda(s_\star - s) + \frac{1}{2\sigma^2}(\|P_{J_\star}^\perp(\mathbf{z})\|^2 - \|P_J^\perp(\mathbf{y})\|^2)\right) \quad (20)$$

with

$$\|P_{J_\star}^\perp(\mathbf{z})\|^2 - \|P_J^\perp(\mathbf{y})\|^2 = 2\mathbf{z}^T(\xi_J - \xi_{J_\star}) + \|\xi_{J_\star}\|^2 - \|\xi_J\|^2. \quad (21)$$

For the inner product on the RHS, note that $\xi_J \in \text{span}(\mathbf{X}_{J \cup J_\star})$ and $\xi_{J_\star} \in \text{span}(\mathbf{X}_{J_\star})$, so that

$$|2\mathbf{z}^T(\xi_J - \xi_{J_\star})| = |2(P_{J \cup J_\star} \mathbf{z})^T(\xi_J - \xi_{J_\star})| \leq 2\|P_{J \cup J_\star} \mathbf{z}\| \|\xi_J - \xi_{J_\star}\|, \quad (22)$$

by Cauchy-Schwarz's inequality.

Lemma 2. For any $c > 0$, with probability at least $1 - p^{-c}$,

$$\|P_J \mathbf{z}\|^2 \leq (20 + 4c)\sigma^2|J| \log p, \quad \forall J \subset [p]. \quad (23)$$

Set $\zeta_J = \sqrt{(20 + 4c)(|J| + s_\star) \log p}$. Using Lemma 2 in (22), from (21) we have

$$\begin{aligned} \|P_{J_\star}^\perp(\mathbf{z})\|^2 - \|P_J^\perp(\mathbf{y})\|^2 &\leq \sigma\zeta_J \|\xi_J - \xi_{J_\star}\| + \|\xi_{J_\star}\|^2 - \|\xi_J\|^2 \\ &\leq \sigma\zeta_J (\|\xi_J\| + \|\xi_{J_\star}\|) + \|\xi_{J_\star}\|^2 - \|\xi_J\|^2 \\ &\leq 4\sigma^2\zeta_J^2 + \frac{3}{2}\|\xi_{J_\star}\|^2 - \frac{1}{2}\|\xi_J\|^2 \\ &\leq 6\sigma^2\zeta_J^2 - \frac{1}{2}\|\xi_J\|^2, \end{aligned} \quad (24)$$

where we used the identity $ab \leq 2a^2 + b^2/2$ in the third inequality, and Lemma 2 to bound $\|\xi_{J_\star}\|^2$ in the last inequality.

We tackle the first part. By definition, $\Pi(\hat{J}_{\text{map}}) \geq \Pi(J_\star)$. Take any J such that $\Pi(J) \geq \Pi(J_\star)$ and let $s = |J|$. Plugging in the bound (24) into (20), and using some crude bounds, we have

$$\begin{aligned} 1 \leq \frac{\Pi(J)}{\Pi(J_\star)} &\leq \exp\left(s_\star \log p + \lambda(s_\star - s) + 3(s + s_\star)(20 + 4c) \log p - \frac{1}{4\sigma^2}\|\xi_J\|^2\right) \\ &\leq \exp\left(s_\star(\lambda + (61 + 12c) \log p) - \frac{1}{4\sigma^2}\|\xi_J\|^2\right), \end{aligned}$$

where we used the fact that $\lambda \geq (62 + 12c) \log p$ in the last inequality. This in turn implies

$$\|\xi_J\|^2 \leq 4\sigma^2 \cdot (\lambda s_\star + (61 + 12c) \log p) \leq 8\sigma^2 \lambda,$$

and the first part of (8) follows from that.

We now turn to the second part. Define $\mathcal{J} = \{J : \|\xi_J\| > \sigma\sqrt{10s_\star\lambda}\}$. We have

$$\begin{aligned} \|\mathbf{X}\hat{\beta}_{\text{mean}} - \mathbf{X}\beta_\star\| &\leq \sum_J \|\xi_J\| \Pi(J) \\ &\leq \sigma\sqrt{10\lambda s_\star} \sum_{J \notin \mathcal{J}} \Pi(J) + \sum_{J \in \mathcal{J}} \|\xi_J\| \frac{\Pi(J)}{\Pi(J_\star)}. \end{aligned} \quad (25)$$

By (20) and (24), we have

$$\begin{aligned}\|\xi_J\| \frac{\Pi(J)}{\Pi(J_\star)} &\leq \|\xi_J\| \frac{\binom{p}{s_\star}}{\binom{p}{s}} \exp\left(\lambda(s_\star - s) + 3\zeta_J^2 - \frac{1}{4\sigma^2}\|\xi_J\|^2\right) \\ &\leq \frac{\sqrt{10}\sigma}{\binom{p}{s}} \exp\left(\lambda(s_\star - s) + s_\star \log p + 3\zeta_J^2 - \frac{1}{5\sigma^2}\|\xi_J\|^2\right),\end{aligned}$$

where we used the fact that $xe^{-x^2} \leq 1/\sqrt{2}$ for all x , and $\binom{p}{s_\star} \leq p^{s_\star}$. Hence, since $\lambda \geq (62+4c)\log p$, we have

$$\begin{aligned}\sum_{J \in \mathcal{J}} \|\xi_J\| \frac{\Pi(J)}{\Pi(J_\star)} &\leq \sum_{s=0}^{\bar{s}} \sum_{J: |J|=s} \frac{\sqrt{10}\sigma}{\binom{p}{s}} \exp(\lambda(s_\star - s) + s_\star \log p + 3\zeta_J^2 - 2\lambda s_\star) \\ &\leq \sqrt{10}\sigma \sum_{s=0}^{\bar{s}} \exp(-(s_\star + s)(\lambda - (61 + 12c)\log p)) \\ &= \sqrt{10}\sigma \cdot 2 \exp(-s_\star(\lambda - (61 + 12c)\log p)) \\ &\leq 2\sqrt{10}\sigma p^{-s_\star}.\end{aligned}\tag{26}$$

The result now follows from

$$\sigma\sqrt{10\lambda s_\star} + 2\sqrt{10}\sigma p^{-s_\star} \leq \sqrt{10}\sigma(\sqrt{\lambda s_\star} + 1) \leq \sigma\sqrt{12\lambda s_\star},$$

since $p \geq 2$ and $s_\star \geq 1$, as well as $\lambda \geq 25$.

6.2 Proof of Proposition 1

Remember (20). We reformulate (21) in the following way

$$\begin{aligned}\|\mathbf{P}_{J_\star}^\perp(\mathbf{z})\|^2 - \|\mathbf{P}_J^\perp(\mathbf{y})\|^2 &= \mathbf{y}^\top (\mathbf{P}_J - \mathbf{P}_{J_\star})\mathbf{y} \\ &= -\|\mathbf{P}_J^\perp \mathbf{X}\beta_\star\|^2 - 2\langle \mathbf{P}_J^\perp \mathbf{X}\beta_\star, \mathbf{z} \rangle + \mathbf{z}^\top (\mathbf{P}_J - \mathbf{P}_{J_\star})\mathbf{z}.\end{aligned}\tag{27}$$

Let $\mathcal{J}_{s,t} = \{J \subset [p] : |J| = s, |J \cap J_\star| = t, J \neq J_\star\}$. We first bound the inner product in (27).

Lemma 3. *For any $c > 0$, with probability at least $1 - p^{-c}$,*

$$\frac{\langle \mathbf{P}_J^\perp \mathbf{X}\beta_\star, \mathbf{z} \rangle^2}{\|\mathbf{P}_J^\perp \mathbf{X}\beta_\star\|^2} \leq (10 + 2c)\sigma^2(s \vee s_\star - t) \log p,\tag{28}$$

for all $J \in \mathcal{J}_{s,t}$ with $t \leq s \wedge s_\star$.

We now bound the quadratic term in (27).

Lemma 4. *For any $c > 0$, with probability at least $1 - p^{-c}$,*

$$\mathbf{z}^\top (\mathbf{P}_J - \mathbf{P}_{J_\star})\mathbf{z} \leq (20 + 4c)\sigma^2(s \vee s_\star - t) \log p,\tag{29}$$

for all $J \in \mathcal{J}_{s,t}$ with $t \leq s \wedge s_\star$.

For a subset $J \subset [p]$, set

$$\gamma_J = \|\mathbf{P}_J^\perp \mathbf{X} \beta_\star\|. \quad (30)$$

Assume that both (28) and (29) hold, which is true with probability at least $1 - 2p^{-c}$. Then, we have that, for all $J \in \mathcal{J}_{s,t}$:

$$\begin{aligned} \mathbf{y}^\top (\mathbf{P}_J - \mathbf{P}_{J_\star}) \mathbf{y} &\leq -\gamma_J^2 + 2\gamma_J \sigma \sqrt{(10 + 2c)(s \vee s_\star - t) \log p} + (20 + 4c)\sigma^2 (s \vee s_\star - t) \log p \\ &\leq (40 + 8c)\sigma^2 (s \vee s_\star - t) \log p - \frac{1}{2}\gamma_J^2 \end{aligned} \quad (31)$$

$$\leq (40 + 8c)\sigma^2 (s \vee s_\star - t) \log p. \quad (32)$$

The first inequality comes from (27), (28) and (29). The identity $2ab \leq a^2 + b^2$, with $a = \gamma_J/\sqrt{2}$ and $b = \sigma \sqrt{(20 + 4c)(s \vee s_\star - t) \log p}$, justifies the second inequality.

Combining (20) and (32), we get

$$\begin{aligned} \sum_{J: |J| > [(1+\varepsilon)s_\star]}^{\bar{s}} \frac{\Pi(J)}{\Pi(J_\star)} &= \sum_{s=[(1+\varepsilon)s_\star]}^{\bar{s}} \sum_{t=0}^{s_\star} \sum_{J \in \mathcal{J}_{s,t}} \frac{\binom{p}{s_\star}}{\binom{p}{s}} \exp \left(\lambda(s_\star - s) + \frac{1}{2\sigma^2} \mathbf{y}^\top (\mathbf{P}_J - \mathbf{P}_{J_\star}) \mathbf{y} \right) \\ &\leq \sum_{s=[(1+\varepsilon)s_\star]}^{\bar{s}} \sum_{t=0}^{s_\star} \frac{\binom{s_\star}{t} \binom{p-s_\star}{s-t} \binom{p}{s_\star}}{\binom{p}{s}} \exp (\lambda(s_\star - s) + (20 + 4c)(s - t) \log p), \end{aligned}$$

where we used the fact that $|\mathcal{J}_{s,t}| = \binom{s_\star}{t} \binom{p-s_\star}{s-t}$ in the last inequality.

For the fraction of binomial coefficients, we have

$$\frac{\binom{s_\star}{t} \binom{p-s_\star}{s-t} \binom{p}{s_\star}}{\binom{p}{s}} = \binom{s}{t} \binom{p-s}{s_\star - t}.$$

We then use the standard bound on the binomial coefficient

$$\begin{aligned} \log \binom{s}{t} + \log \binom{p-s}{s_\star - t} &\leq (s-t) \log (es/(s-t)) + (s_\star - t) \log (e(p-s)/(s_\star - t)) \\ &\leq 3(s \vee s_\star - t) \log p. \end{aligned} \quad (33)$$

Hence, we have so far that

$$\sum_{J: |J| > [(1+\varepsilon)s_\star]}^{\bar{s}} \frac{\Pi(J)}{\Pi(J_\star)} \leq \sum_{s=0}^{\bar{s}} \sum_{t=0}^{s_\star} \exp (A_{s,t}), \quad (34)$$

where

$$A_{s,t} := \omega(s-t) \log p + \lambda(s_\star - s), \quad \omega := 23 + 4c.$$

Some simple algebra yields

$$\begin{aligned} \sum_{s \geq [(1+\varepsilon)s_\star]}^{\bar{s}} \sum_{t=0}^{s_\star} \exp (A_{s,t}) &\leq \sum_{s \geq (1+\varepsilon)s_\star} e^{-(\lambda - \omega \log p)(s-s_\star)} \sum_{t=0}^{s_\star} e^{(s_\star - t)\omega \log p} \\ &\leq \frac{e^{-(\lambda - \omega \log p)\varepsilon s_\star}}{1 - e^{-\lambda + \omega \log p}} \cdot \frac{e^{(s_\star + 1)\omega \log p}}{e^{\omega \log p} - 1} \end{aligned} \quad (35)$$

$$\leq \frac{p^{-c}}{(1 - p^{-\omega})(1 - p^{-c})}, \quad (36)$$

where we used the fact that $p^\omega \geq 2$, because $p \geq 2$, and also $-(\lambda - \omega \log p)\varepsilon s_\star + s_\star \omega \log p \leq -c \log p$, because of (9). This shows that

$$\begin{aligned} \Pi(J : |J| > [(1 + \varepsilon)s_\star]) &\leq \frac{p^{-c}}{(1 - p^{-\omega})(1 - p^{-c})} \Pi(J_\star) \\ &\leq \frac{p^{-c}}{(1 - p^{-c})^2} \Pi(J_\star), \end{aligned}$$

using the fact that $\omega \geq c$. From this, and the fact that $p^{-c} \leq 1/2$, we conclude the proof.

6.3 Proof of Theorem 2

Let $\nu = \nu_{(2+\varepsilon)s_\star}$ for short. The proof of this result is identical to that of Proposition 1 up to (31). We now need a lower bound on γ_J . For this, we use the following irrepresentability result.

Lemma 5. *Let $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2]$, with smallest singular value δ , and let \mathbf{P}_2 denote the orthogonal projection onto \mathbf{X}_2 . Then for any β_1 ,*

$$\|(\mathbf{I} - \mathbf{P}_2)\mathbf{X}_1\beta_1\| \geq \delta\|\beta_1\|.$$

Note that for any $J \in \mathcal{J}_{s,t}$ with $s - t \leq (1 + \varepsilon)s_\star$, the smallest singular value of $[\mathbf{X}_{J_\star} \mathbf{X}_{J \setminus J_\star}]$ is bounded from below by $\sqrt{n}\nu$; by Lemma 5, this implies that

$$\gamma_J = \|(\mathbf{I} - \mathbf{P}_J)(\mathbf{X}_{J_\star}\beta_\star)\| = \|(\mathbf{I} - \mathbf{P}_J)(\mathbf{X}_{J_\star \setminus J}\beta_{J_\star \setminus J}^*)\| \geq \sqrt{n}\nu\|\beta_{J_\star \setminus J}^*\|.$$

Hence,

$$\gamma_J \geq \rho\nu\sqrt{n(s_\star - t)}, \quad \forall J \in \mathcal{J}_{s,t}, \text{ such that } 0 \leq t \leq s_\star \wedge s \text{ and } s \leq t + (1 + \varepsilon)s_\star, \quad (37)$$

where we recall that ρ is defined in (12).

In view of (31) and (37) we have, with probability at least $1 - 2p^{-c}$, for all $J \in \mathcal{J}_{s,t}$

$$\begin{aligned} \mathbf{y}^\top(\mathbf{P}_J - \mathbf{P}_{J_\star})\mathbf{y} &\leq (40 + 8c)\sigma^2(s \vee s_\star - t) \log p - \frac{1}{2}\gamma_J^2 \\ &\leq (40 + 8c)\sigma^2(s \vee s_\star - t) \log p - \frac{1}{2}\rho^2\nu^2n(s_\star - t)\mathbb{I}_{\{s \leq t + (1 + \varepsilon)s_\star\}}. \end{aligned} \quad (38)$$

Next, we have

$$\frac{1}{\Pi(J_\star)} = \sum_{J : |J| > [(1 + \varepsilon)s_\star]} \frac{\Pi(J)}{\Pi(J_\star)} + \sum_{J : |J| \leq [(1 + \varepsilon)s_\star]} \frac{\Pi(J)}{\Pi(J_\star)}. \quad (39)$$

The first sum in the right-hand side was already bounded in Proposition 1. We concentrate on the second sum.

Combining (20) and (38), we get

$$\begin{aligned} \sum_{J : |J| \leq [(1 + \varepsilon)s_\star]} \frac{\Pi(J)}{\Pi(J_\star)} &= \sum_{s=0}^{[(1 + \varepsilon)s_\star]} \sum_{t=0}^{s \wedge s_\star} \sum_{J \in \mathcal{J}_{s,t}} \frac{\binom{p}{s_\star}}{\binom{p}{s}} \exp\left(\lambda(s_\star - s) + \frac{1}{2\sigma^2}\mathbf{y}^\top(\mathbf{P}_J - \mathbf{P}_{J_\star})\mathbf{y}\right) \\ &\leq \sum_{s=0}^{[(1 + \varepsilon)s_\star]} \sum_{t=0}^{s \wedge s_\star} \frac{\binom{s_\star}{t} \binom{p - s_\star}{s - t} \binom{p}{s}}{\binom{p}{s}} \exp\left(\lambda(s_\star - s) + (20 + 4c)(s \vee s_\star - t) \log p - \eta_{s,t}\right), \\ &\leq \sum_{s=0}^{[(1 + \varepsilon)s_\star]} \sum_{t=0}^{s \wedge s_\star} \binom{s}{t} \binom{p - s}{s_\star - t} \exp\left(\lambda(s_\star - s) + (20 + 4c)(s \vee s_\star - t) \log p - \eta_{s,t}\right), \end{aligned}$$

where $\eta_{s,t} := \frac{1}{4\sigma^2} \rho^2 \nu^2 n(s_\star - t) \mathbb{I}_{\{s \leq t + [(1+\varepsilon)s_\star]\}}$.

Next, we use again (33) to get

$$\sum_{J: |J| \leq [(1+\varepsilon)s_\star]} \frac{\Pi(J)}{\Pi(J_\star)} \leq \sum_{s=0}^{[(1+\varepsilon)s_\star]} \sum_{t=0}^{s \wedge s_\star} \exp(A_{s,t}), \quad (40)$$

where

$$A_{s,t} := \omega(s \vee s_\star - t) \log p + \lambda(s_\star - s) - \eta_{s,t}, \quad \omega := 23 + 4c.$$

Let $\alpha = \frac{\nu^2 n \rho^2}{4\sigma^2} - \omega \log p$, and note that $\alpha \geq 2\lambda \geq \lambda + c \log p$ by (9) and (12).

When $s \leq s_\star$, we have $A_{s,t} = -\alpha(s-t) - (\alpha - \lambda)(s_\star - s)$, so that

$$\begin{aligned} \sum_{s=0}^{s_\star} \sum_{t=0}^s \exp(A_{s,t}) &\leq \sum_{s=1}^{s_\star} e^{-(s_\star - s)c \log p} \sum_{t=0}^s e^{-\alpha(s-t)} \\ &\leq \frac{1}{(1 - e^{-\alpha})(1 - p^{-c})}. \end{aligned} \quad (41)$$

When $s_\star < s \leq (1+\varepsilon)s_\star$, we have $A_{s,t} = -\alpha(s_\star - t) - (\lambda - \omega \log p)(s - s_\star)$, with $\lambda \geq \omega \log p + c \log p$, leading to

$$\begin{aligned} \sum_{s=s_\star+1}^{[(1+\varepsilon)s_\star]} \sum_{t=0}^{s_\star} \exp(A_{s,t}) &\leq \sum_{s=s_\star+1}^{\infty} e^{-(s-s_\star)c \log p} \sum_{t=0}^{s_\star} e^{-\alpha(s_\star - t)} \\ &\leq \frac{p^{-c}}{(1 - e^{-\alpha})(1 - p^{-c})}. \end{aligned} \quad (42)$$

Combining (36) with (39)-(42), we conclude that

$$\begin{aligned} \frac{1}{\Pi(J_\star)} &\leq \frac{1}{(1 - e^{-\alpha})(1 - p^{-c})} + \frac{p^{-c}}{(1 - e^{-\alpha})(1 - p^{-c})} + \frac{p^{-c}}{(1 - p^{-\omega})(1 - p^{-c})} \\ &\leq \frac{1 + 2p^{-c}}{(1 - p^{-c})^2}, \end{aligned}$$

using the fact that $\alpha \geq \omega \geq c$. From this, we get

$$\Pi(J_\star) \geq (1 - p^{-c})^2 (1 - 2p^{-c}) \geq (1 - 2p^{-c})^2 \geq 1 - 4p^{-c}.$$

This concludes the proof of Theorem 2. We note that the proof of (13) is virtually identical.

6.4 Proof of Theorem 3

When (9) is satisfied with $\varepsilon \leq 1/2$, then λ satisfies both the conditions of Proposition 1 and Theorem 1. Hence, with probability at least $1 - 2p^{-c} - p^{-c} = 1 - 3p^{-c}$, we have both that $|\hat{J}_{\text{map}}| \leq (1+\varepsilon)s_\star$ and (8). Hence, the support of $\hat{\beta}_{\text{map}} - \beta_\star$ is of size at most $(1+\varepsilon)s_\star + s_\star = (2+\varepsilon)s_\star$, and we have

$$\|\hat{\beta}_{\text{map}} - \beta_\star\| \leq \frac{1}{\nu_{(2+\varepsilon)s_\star}} \|\mathbf{X}(\hat{\beta}_{\text{map}} - \beta_\star)\|,$$

with

$$\|\mathbf{X}(\hat{\beta}_{\text{map}} - \beta_\star)\| = \|\mathbf{X}\hat{\beta}_{\text{map}} - \mathbf{X}\beta_\star\| \leq \sigma \sqrt{8s_\star \lambda},$$

and the result follows.

6.5 Proof of Theorem 4

For $r > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\|\widehat{\beta}_{\text{map}} - \beta_{\star}\|_{\infty} > r\right) &\leq \mathbb{P}\left(\|\widehat{\beta}_{J_{\star}} - \beta_{\star}\|_{\infty} > r, \widehat{J}_{\text{map}} = J_{\star}\right) + \mathbb{P}\left(\|\widehat{\beta}_{\text{map}} - \beta_{\star}\|_{\infty} > r, \widehat{J}_{\text{map}} \neq J_{\star}\right) \\ &\leq \mathbb{P}\left(\|\widehat{\beta}_{J_{\star}} - \beta_{\star}\|_{\infty} > r\right) + \mathbb{P}\left(\widehat{J}_{\text{map}} \neq J_{\star}\right).\end{aligned}$$

By Theorem 2, $\widehat{J}_{\text{map}} = J_{\star}$ with probability at least $1 - 2p^{-c}$, so that the second term on the RHS is bounded by $2p^{-c}$.

Next, we know that $\widehat{\beta}_{J_{\star}} \sim N(\beta_{\star}, \sigma^2 \frac{1}{n} \Psi_{\star}^{-1})$ with $\Psi_{\star} := \frac{1}{n} \mathbf{X}_{J_{\star}}^{\top} \mathbf{X}_{J_{\star}}$, and in particular, $\widehat{\beta}_{J_{\star},j} - \beta_{\star,j} \sim \mathcal{N}(0, \sigma^2 \tau_j^2/n)$, where τ_j^2 is the j th diagonal entry of Ψ_{\star}^{-1} . This matrix being positive semi-definite, its diagonal terms are all bounded from above by its largest eigenvalue, which is the inverse of the smallest eigenvalue of Ψ_{\star} , which in turn is larger than $\nu_{s_{\star}}^2$. Hence, $\text{Var}(\widehat{\beta}_{J_{\star},j}) \leq \sigma^2/(n\nu_{s_{\star}}^2)$ for all $j \in J_{\star}$, so that a standard tail bound on the normal distribution and the union bound give

$$\mathbb{P}\left(\|\widehat{\beta}_{J_{\star}} - \beta_{\star}\|_{\infty} > r\right) \leq s_{\star} \exp\left(-\frac{n\nu_{s_{\star}}^2 r^2}{2\sigma^2}\right). \quad (43)$$

Taking $r = \sigma\sqrt{2(c+1)\log(p)/(n\nu_{s_{\star}}^2)}$ bounds this by p^{-c} , and the desired result follows.

6.6 Proof of Theorem 5

We have

$$\begin{aligned}\|\widehat{\beta}_{\text{map}} - \beta_{\star}\|_{\infty} &\leq \sum_J \|\widehat{\beta}_J - \beta_{\star}\|_{\infty} \Pi(J) \\ &\leq \|\widehat{\beta}_{J_{\star}} - \beta_{\star}\|_{\infty} \Pi(J_{\star}) + \sum_{J \neq J_{\star}} \|\widehat{\beta}_J - \beta_{\star}\|_{\infty} \Pi(J) \\ &\leq \|\widehat{\beta}_{J_{\star}} - \beta_{\star}\|_{\infty} + \sum_{J \neq J_{\star}} \|\widehat{\beta}_J - \beta_{\star}\|_{\infty} \Pi(J).\end{aligned} \quad (44)$$

For any $c > 0$, we have with probability at least $1 - p^{-c}$, for any $J \subset [p]$ with $\nu_J > 0$, that

$$\begin{aligned}\|\widehat{\beta}_J\|_{\infty} &\leq \sqrt{|J|} \|\widehat{\beta}_J\| \\ &\leq \frac{\sqrt{|J|}}{\sqrt{n\nu_J}} \|\mathbf{X} \widehat{\beta}_J\| \\ &\leq \frac{\sqrt{|J|}}{\sqrt{n\nu_J}} \left[\|\mathbf{P}_J(z)\| + \|\mathbf{P}_J^{\perp}(\mathbf{X} \beta_{\star})\| \right] \\ &\leq \frac{\sqrt{|J|}}{\sqrt{n\nu_J}} \left[\sigma \sqrt{(20+4c)|J| \log p} + \|\mathbf{X} \beta_{\star}\| \right],\end{aligned}$$

where we have used Cauchy-Schwarz's inequality in the first line and (23) in the last line.

We now assume that $\nu_{\overline{s}} > 0$, which implies that $\nu_J > 0$ for any $J \subset [p]$ with $|J| \leq \overline{s}$. Combining the previous display with (43) and (44) and a union bound argument, we get with probability at least $1 - 2p^{-c}$,

$$\begin{aligned}\|\widehat{\beta}_{\text{map}} - \beta_{\star}\|_{\infty} &\leq \sigma \sqrt{\frac{2(c+1)\log p}{n\nu_{s_{\star}}}} \\ &\quad + \sum_{J \neq J_{\star}} \left[\frac{\sigma|J|}{\nu_{\overline{s}}} \sqrt{(20+4c)\log p} + \frac{\sqrt{|J|}}{\sqrt{n\nu_{\overline{s}}}} \|\mathbf{X} \beta_{\star}\| + \|\beta_{\star}\|_{\infty} \right] \Pi(J).\end{aligned}$$

Next, we combine the above display with (13) and a union bound argument to get with probability at least $1 - 4p^{-c}$ that

$$\|\widehat{\beta}_{\text{map}} - \beta_{\star}\|_{\infty} \leq \sigma \sqrt{\frac{2(c+1)\log p}{n\nu_{s_{\star}}}} + \frac{4p^{-c}}{\nu_{\bar{s}}} \left[\sigma \sqrt{(20+4c)\frac{\log p}{n}} + \frac{\|\mathbf{X}\beta_{\star}\|}{\sqrt{n}} + \nu_{\bar{s}}\|\beta_{\star}\|_{\infty} \right].$$

Note that the same reasoning applied to $\widetilde{\beta}$ yields the same l_{∞} -norm estimation bound with $\nu_{\bar{s}}$ replaced by ν_{\min} .

We now assume that $\nu_{s_{\star}+\bar{s}} > 0$. Then, for any $J \subset [p]$ with $|J| \leq \bar{s}$, we have

$$\|\widehat{\beta}_J - \beta_{\star}\|_{\infty} \leq \|\widehat{\beta}_J - \beta_{\star}\| \leq \frac{\|\mathbf{X}\widehat{\beta}_J - \mathbf{X}\beta_{\star}\|}{\sqrt{n\nu_{s_{\star}+\bar{s}}}}.$$

Combining this last inequality with (44), we get

$$\begin{aligned} \|\widehat{\beta}_{\text{map}} - \beta_{\star}\|_{\infty} &\leq \|\widehat{\beta}_{J_{\star}} - \beta_{\star}\|_{\infty} + \frac{1}{\sqrt{n\nu_{s_{\star}+\bar{s}}}} \sum_{J \notin \mathcal{J}, J \neq J_{\star}} \|\xi_J\| \Pi(J) + \frac{1}{\sqrt{n\nu_{s_{\star}+\bar{s}}}} \sum_{J \in \mathcal{J}} \|\xi_J\| \Pi(J) \\ &\leq \|\widehat{\beta}_{J_{\star}} - \beta_{\star}\|_{\infty} + \frac{\sigma\sqrt{10s_{\star}}}{\sqrt{n\nu_{s_{\star}+\bar{s}}}} \Pi(\mathcal{J}^c \setminus J_{\star}) + \frac{1}{\sqrt{n\nu_{s_{\star}+\bar{s}}}} \sum_{J \in \mathcal{J}} \|\xi_J\| \Pi(J), \end{aligned}$$

where we recall that $\xi_J = \mathbf{X}\widehat{\beta}_J - \mathbf{X}\beta_{\star}$ and $\mathcal{J} = \{J \subset [p] : \|\xi_J\| > \sigma\sqrt{10s_{\star}\lambda}\}$. In view of Theorem 2, we have with probability at least $1 - 2p^{-c}$ that

$$\Pi(\mathcal{J}^c \setminus J_{\star}) \leq 1 - \Pi(J_{\star}) \leq 4p^{-c};$$

and in view of (26),

$$\sum_{J \in \mathcal{J}} \|\xi_J\| \Pi(J) \leq 2\sqrt{10}\sigma p^{-s_{\star}}.$$

Combining the three last displays with (43), we get the result.

6.7 Proofs of auxiliary results

Lemma 2 is a special case of Lemma 4 where $J_{\star} = \emptyset$, and we prove Lemma 4 below.

6.7.1 Proof of Lemma 3

First, note that $u_J := \langle \mathbf{P}_J^{\perp} \mathbf{X}\beta_{\star}, \mathbf{z} \rangle \sim \mathcal{N}(0, \sigma^2 \gamma_J^2)$, where γ_J is defined in (30), so that $v_J := u_J/(\sigma\gamma_J) \sim \mathcal{N}(0, 1)$. By the union bound and a standard tail bound on the normal distribution, for $a > 0$, we have

$$\mathbb{P}\left(\max_{J \in \mathcal{J}_{s,t}} v_J^2 > a^2\right) \leq \binom{s_{\star}}{t} \binom{p-s_{\star}}{s-t} \exp(-a^2/2).$$

As in (33), we have

$$\begin{aligned} \log \binom{s_{\star}}{t} + \log \binom{p-s_{\star}}{s-t} &\leq (s_{\star} - t) \log(es_{\star}) + (s - t) \log(ep) \\ &\leq 3(s \vee s_{\star} - t) \log p. \end{aligned} \tag{45}$$

Hence,

$$\mathbb{P} \left(\max_{J \in \mathcal{J}_{s,t}} v_J^2 > (10 + 2c)(s \vee s_\star - t) \log p \right) \leq \exp \left(-(2 + c)(s \vee s_\star - t) \log p \right) \leq p^{-(2+c)},$$

since $s \vee s_\star - t = 0$ would imply $J = J_\star$. We then apply the union bound again,

$$\mathbb{P} \left(\max_{s,t} \max_{J \in \mathcal{J}_{s,t}} \frac{v_J^2}{s \vee s_\star - t} > (10 + 2c)\sigma^2 \log p \right) \leq \bar{s}(s \wedge s_\star + 1)p^{-(2+c)} \leq p^{-c},$$

which the result we wanted.

6.7.2 Proof of Lemma 4

Fix $J \in \mathcal{J}_{s,t}$. First, we notice that

$$\mathbf{z}^\top (\mathbf{P}_J - \mathbf{P}_{J_\star}) \mathbf{z} = \mathbf{z}^\top (\mathbf{P}_J - \mathbf{P}_{J \cap J_\star}) \mathbf{z} - \mathbf{z}^\top (\mathbf{P}_{J_\star} - \mathbf{P}_{J \cap J_\star}) \mathbf{z} \leq \mathbf{z}^\top (\mathbf{P}_J - \mathbf{P}_{J \cap J_\star}) \mathbf{z},$$

since $\mathbf{P}_{J_\star} - \mathbf{P}_{J \cap J_\star}$ is an orthogonal projection, and therefore positive semidefinite. And $\mathbf{Q}_J := \mathbf{P}_J - \mathbf{P}_{J \cap J_\star}$ is also an orthogonal projection, of rank $s - t$, so that $\|\mathbf{Q}_J \mathbf{z}\|^2 \sim \sigma^2 \chi_{s-t}^2$. Chernoff's Bound applied to the chi-square distribution yields

$$\log \mathbb{P}(\chi_m^2 > a) \leq -\frac{m}{2}(a/m - 1 - \log(a/m)) \leq -\frac{a}{4}, \quad \forall a \geq 2m.$$

The union bound and (45), and this tail bound, yields

$$\mathbb{P} \left(\max_{J \in \mathcal{J}_{s,t}} \|\mathbf{Q}_J \mathbf{z}\|^2 > (20 + 4c)\sigma^2(s \vee s_\star - t) \log p \right) \leq \exp \left(-(2 + c)(s \vee s_\star - t) \log p \right).$$

The rest of the proof is exactly the same as that of Lemma 3.

6.8 An irrepresentability result

We have

$$\begin{aligned} \|(I - \mathbf{P}_2)\mathbf{X}_1\boldsymbol{\beta}_1\|^2 &= \min_{\boldsymbol{\beta}_2} \|\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2\|^2 \\ &= \min_{\boldsymbol{\beta}_2} \boldsymbol{\beta} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} \\ &\geq \min_{\boldsymbol{\beta}_2} \delta^2 \|\boldsymbol{\beta}\|^2 \\ &= \delta^2 \|\boldsymbol{\beta}_1\|^2, \end{aligned}$$

where $\boldsymbol{\beta} := (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, implying $\|\boldsymbol{\beta}\|^2 = \|\boldsymbol{\beta}_1\|^2 + \|\boldsymbol{\beta}_2\|^2$.

Acknowledgements

We would like to thank Pierre Alquier, Philippe Rigollet and Alexander Tsybakov for sharing their R codes with us. This work was partially supported by XXX.

References

- Alquier, P. and K. Lounici (2011). Pac-bayesian theorems for sparse regression estimation with exponential weights. *Electronic Journal of Statistics* 5, 127–145. Arxiv:1009.2707.
- Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, New York, NY, USA, pp. 33–40. ACM.
- Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics* 37(4), 1705–1732.
- Bunea, F. (2008). Consistent selection via the Lasso for high dimensional approximating regression models. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, Volume 3 of *Inst. Math. Stat. Collect.*, pp. 122–137. Beachwood, OH: Inst. Math. Statist.
- Bunea, F., A. Tsybakov, and M. Wegkamp (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* 1, 169–194.
- Cai, T. and L. Wang (2011). Orthogonal matching pursuit for sparse signal recovery with noise. 57(7), 4680–4688.
- Candès, E. and M. Davenport (2011). How well can we estimate a sparse vector? *Arxiv preprint arXiv:1104.5246*.
- Candès, E. and Y. Plan (2009). Near-ideal model selection by ℓ_1 minimization. 37(5A), 2145–2177.
- Candès, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n . *Ann. Statist.* 35(6), 2313–2351.
- Catoni, O. (2004). *Statistical learning theory and stochastic optimization*, Volume 1851 of *Lecture Notes in Mathematics*. Berlin: Springer-Verlag. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.
- Chen, J. and Z. Chen (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 95(3), 759–771.
- Chipman, H., E. I. George, and R. E. McCulloch (2001). The practical implementation of Bayesian model selection. In *Model selection*, Volume 38 of *IMS Lecture Notes Monogr. Ser.*, pp. 65–134. Beachwood, OH: Inst. Math. Statist. With discussion by M. Clyde, Dean P. Foster, and Robert A. Stine, and a rejoinder by the authors.
- Dalalyan, A. and J. Salmon (2011). Optimal aggregation of affine estimators. In *Proceedings of the 24th annual conference on Computational Learning Theory*, Budapest (Hungary).
- Dalalyan, A. and A. Tsybakov (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, Volume 4539 of *Lecture Notes in Comput. Sci.*, pp. 97–111. Berlin: Springer.
- Fan, J. and R. Li (2001). Sure independence screening for ultrahigh dimensional feature space. *J. Amer. Statist. Assoc.* 96, 1348–1360.
- Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5), 849–911.
- Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* 57(8), 5467–5484.
- Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* 32(3), 928–961.
- Gautier, E. and A. Tsybakov (2011, October). High-dimensional instrumental variables regression and confidence sets. Technical report, Arxiv preprint 1105.2454v3.

- Giraud, C. (2007). Mixing least-squares estimators when the variance is unknown. Technical report, Arxiv preprint arXiv:0711.0372.
- Ji, P. and J. Jin (2010). Ups delivers optimal phase diagram in high dimensional variable selection. *Arxiv preprint arXiv:1010.5028*.
- Jin, J., C. Zhang, and Q. Zhang (2012). Optimality of graphlet screening in high dimensional variable selection. *Arxiv preprint arXiv:1204.6452*.
- Juditsky, A., P. Rigollet, and A. B. Tsybakov (2008). Learning by mirror averaging. *Ann. Statist.* 36(5), 2183–2206.
- Leung, G. and A. Barron (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* 52(8), 3396–3410.
- Lounici, K. (2007). Generalized mirror averaging and D -convex aggregation. *Math. Methods Statist.* 16(3), 246–259.
- Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics* 2, 90–102.
- Lounici, K. (2009). *Statistical Estimation in High-Dimension, Sparsity Oracle Inequalities*. Ph. D. thesis, University Paris Diderot - Paris 7.
- Lounici, K., M. Pontil, A. Tsybakov, and S. van de Geer (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* 39(4), 2164–2204.
- Meinshausen, N., P. Bühlmann, and E. Züchrich (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics* 34, 1436–1462.
- Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* 37(1), 246–270.
- Raskutti, G., M. Wainwright, and B. Yu (2009). Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *Arxiv preprint arXiv:0910.2042*.
- Rigollet, P. and A. Tsybakov (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* 39(2), 731–771.
- Rigollet, P. and A. Tsybakov (2012). Sparse estimation by exponential weighting. Technical report, Arxiv preprint.
- Robert, C. and G. Casella (2004). *Monte Carlo statistical methods*. Computer Science and Scientific Computing. Springer Verlag.
- Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* 7(2), 221–264. With comments and a rejoinder by the author.
- Stewart, G. W. and J. G. Sun (1990). *Matrix perturbation theory*. Computer Science and Scientific Computing. Boston, MA: Academic Press Inc.
- Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. Available from <http://arxiv.org/abs/1011.3027>.
- Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* 6, 38–90.
- Wainwright, M. (2006). Sharp thresholds for noisy and high-dimensional recovery of sparsity using l_1 -constrained quadratic programming.
- Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* 10(8), 25–47.
- Zhang, C.-H. (2007). Information-theoretic optimality of variable selection with concave penalty. Technical report, Dept. Statistics, Rutgers Univ.
- Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* 38(2), 894–942.
- Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* 7, 2541–2563.